Life Science Research Practice: A Revolution in Progress

Jim Myers (Jim.Myers@pnl.gov), Ron Taylor (Ron.Taylor@pnl.gov), Yuri Gorby (Yuri.Gorby@pnl.gov), Debbie Gracio (Debbie.Gracio@pnl.gov), Pacific Northwest National Laboratory

DRAFT requirements white paper in support of the DOE National Collaboratory Program - January, 2005 – please send comments to Jim.Myers@pnl.gov

Life Sciences research is undergoing a transformation from being qualitative and descriptive to being quantitative and predictive. The availability of high-throughput, data-intensive “omics” technologies, such as genomics, transcriptomics, proteomics, and metabolomics, together with advances in computationally intensive simulation and the availability of community data repositories, is revolutionizing the way biological research is conducted, creating, for the first time, an opportunity to address exciting biological questions such as the following [1]:

- What biochemical pathways control a plant’s ability to create biomass, or a microbe’s ability to produce hydrogen?
- Can we identify natural populations of microbes that degrade or immobilize contaminants such as hydrocarbons or metals?
- What cellular repair mechanisms are employed by bacteria that live in environments with ionizing radiation?
- What communities of microbes are most effective in taking up excess carbon from the atmosphere?

The key challenges in achieving the systems-level understanding necessary to address such questions are coordination and integration – of expertise, experimental and computational resources, and myriad types of data – from multiple independent groups and multiple locations. Due to the complexity and scale of life sciences research, collaboration is required across distributed groups as well as within and across communities. Life scientists have been at the forefront in developing, evaluating, and adopting collaborative technologies ranging from remotely controlled instruments [2,3], collaborative portals [4,5], and, most successfully, community databases and computational services [6,7]. More recent projects are exploring the use of cross-scale data integration [8] and the use of semantic and grid technologies to simplify data integration and workflow [9-11].

In the next five years, data-centric collaboration in the life sciences will expand from its roots in genomics and bioinformatics and become a standard component of next-generation biological research. As it does so, collaboration and cyberinfrastructure technologies will need to scale by orders of magnitude in terms of the number of users and groups employing them, the amount of data accessible, and in the heterogeneity and complexity of information being assembled. In essence, systems-oriented research will aim to produce "big science" results by integrating the effort of thousands of independent research programs developing model organisms, instruments, theories, analysis techniques, and models. Research will be planned by distributed teams targeting multiple outcomes, encompass data from myriad techniques from distributed data repositories, invoke analysis services and multi-scale models hosted at remote compute centers, and be richly documented with provenance and other metadata to allow reanalysis and reuse by others.

High-throughput experiments and simulations already are generating vast amounts of complex data. For example, high-end Fourier transform ion cyclotron resonance (FTICR) mass spectrometers generate 20 GB per sample, and facilities are already generating 250+ GB per day. High-throughput proteomics facilities such as those planned as part of the DOE Genomics:GTL program will be able to analyze hundreds of samples per day, providing hundreds of petabytes of data per year within the next decade. These data need to be analyzed and interpreted within specific research efforts and then documented and shared in order to create knowledge bases. Similarly, biomolecular simulations that relate structure and function of biological systems will be generating hundreds of gigabytes for each simulation run. All this information, and tools to support meaningful integration, visualization, and comparison of results from different techniques, need to be shared, annotated, archived, and made accessible to the general biological community. The scale of these tasks will require new mechanisms throughout the research lifecycle to, for example, capture rich data and model descriptions; document data accuracy, quality, and uncertainty; integrate heterogeneous information from independent sources; and perform data mining and visualization of high-dimensional information. These technologies will need to be assembled into advanced environments that support high-level tasks such as discovering existing resources, specifying and executing complex workflows and queries, and documenting research while hiding the complex mechanics involved. Further, while the requirements to scale these efforts to support the entire life sciences community will necessity increased reliance on self-describing data and services, personal interactions within and between distributed groups will also be important for planning, discussion, decision making, and troubleshooting.

The requirements noted here represent significant challenges. However, the existing knowledge and technology base across data and information management, distributed computing, collaboratories, and semantic information processing strongly suggest that the requirements can be met. Success in developing and deploying next-generation data-centric collaboration capabilities in the life sciences will be critical to maximizing the value of biological research and in improving our lives.

Online References:

1. Biological questions and data management discussion adapted from input provided for the DOE Data Management Workshops 2004 (http://www-conf.slac.stanford.edu/dmw2004/) Report, in preparation
2. EMSL Virtual NMR Facility, http://collaboratory.emsl.pnl.gov/virtual/EMSLVNMRF.html
3. Telescience^{TM}, A Collaborative Environment for Telemicroscopy^{TM} and Remote Science, https://telescience.ucsd.edu/

4. BioCoRE - A Biological Collaborative Research Environment, http://www.ks.uiuc.edu/Research/biocore/
5. BioSPICE Community Web Site :: Biology in silicio, https://community.biospice.org/
6. The RCSB Protein Data Bank, http://www.rcsb.org/pdb/
7. Basic Local Alignment Search Tool, http://www.ncbi.nlm.nih.gov/blast/
8. Biomedical Informatics Research network – Home Page, http://www.nbirn.net/
9. MicroArray and Gene Expression, http://www.mged.org/Workgroups/MAGE/mage.html
10. myGrid, http://www.mygrid.org.uk/
11. cancer Biomedical Informatics Grid, http://cabig.nci.nih.gov/